ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐÀM PHƯƠNG TÙNG

PHÂN LOẠI BÌNH LUẬN CỦA KHÁCH HÀNG TRÊN MẠNG XÃ HỘI DỰA TRÊN KỸ THUẬT MÁY HỌC

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN 2020

LỜI CAM ĐOAN

Tôi xin cam đoan kết quả đạt được trong luận văn là sản phẩm của cá nhân dưới sự hướng dẫn khoa học của TS. Nguyễn Văn Tảo. Trong toàn bộ nội dung luận văn, những nội dung được trình bày là của cá nhân hoặc tổng hợp từ nhiều nguồn tài liệu khác nhau. Tất cả các tài liệu tham khảo đó đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Thái Nguyên, tháng năm 2020

Tác giả

Đàm Phương Tùng

LỜI CẢM ƠN

Học viên xin bày tỏ lời cảm ơn chân thành tới tập thể các thầy cô giáo Viện công nghệ thông tin, các thầy cô giáo Trường Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên đã mang lại cho học viên kiến thức vô cùng quý giá và bổ ích trong suốt quá trình học tập chương trình cao học tại trường. Đặc biệt học viên xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo TS.NGUYỄN VĂN TẢO đã định hướng khoa học và đưa ra những góp ý, gợi ý, chỉnh sửa quý báu, quan tâm, tạo điều kiện thuận lợi trong quá trình nghiên cứu hoàn thành luận văn này.

Cuối cùng, học viên xin chân thành cảm ơn các bạn bè đồng nghiệp, gia đình và người thân đã quan tâm, giúp đỡ và chia sẻ với học viên trong suốt quá trình học tập.

Do thời gian và kiến thức có hạn nên luận văn chắc không tránh khỏi những thiếu sót nhất định. Học viên rất mong nhận được những sự góp ý quý báu của thầy cô và các bạn.

Thái Nguyên, tháng năm 2020

Tác giả

Đàm Phương Tùng

MỤC LỤC

DANH MỤC HÌNH ẢNH

LỜI MỞ ĐẦU

Trong thời buổi kinh tế thị trường ngày hôm nay, một doanh nghiệp muốn tồn tại và phát triển thì cần phải khai thác và thu thập được các bình luận phản hồi của người dùng về sản phẩm hay dịch vụ của mình từ đó đưa ra những định hướng và điều chỉnh về hoạt động sản xuất kinh doanh phù hợp hơn.

Cùng với sự ra đời của internet, sự xuất hiện và phát triển không ngừng của lĩnh vực thương mại điện tử khiến cho việc xúc tiến các hoạt động kinh doanh, buôn bán, quảng bá sản phẩm, dịch vụ diễn ra trên khắp các kênh thông tin xã hội đặc biệt là trên mạng internet. Điều này vô hình dung tạo nên cầu nối giữa người dùng và nhà cung cấp, và từ cầu nối này người dùng có thể đưa ra bình luận của họ đối với sản phẩm hay dịch vụ mà nhà cung cấp mang lại.

Như chúng ta đã biết ngày nay mọi thông tin đều được đưa lên các trang mạng xã hội dưới dạng các posts và rất nhiều người dùng để lại các các nhận xét của mình về các posts này dưới dạng các comments, ta nhận thấy đây là kho thông tin khổng lồ mà từ đó nếu chúng ta có thể khai phá và trích rút tất cả các comments của người dùng, sau đó phân tích và phân loại dữ liệu ấy, chúng ta có thể thu được các kết quả khảo sát cần thiết phục vụ cho hoạt động sản xuất kinh doanh. Việc phân loại bình luận khách hàng về nhiều lĩnh vực, giúp doanh nghiệp có cách quản lý tốt hơn, đưa ra những sáng kiến mới giúp doanh nghiệp mình phát triển.

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ BÀI TOÁN PHÂN LOẠI BÌNH LUẬN KHÁCH HÀNG

1.1. Tổng quan về khai phá dữ liệu

KPDL là một trong những thuật ngữ mới xuất hiện đầu thế kỷ 21, nó là hệ quả của sự bùng nổ Internet đạt tới đỉnh điểm. Theo một công bố của Intel vào tháng 9 năm 2013, cứ 11 giây trôi qua lại có thêm 1 Petabybe dữ liệu, nó tương đương với một video chất lượng HD dài 13 năm.

KPDL đã và đang được ứng dụng rộng rãi trong rất nhiều lĩnh vực và hiện nay đã có rất nhiều công cụ thương mại và phi thương mại triển khai các nhiệm vụ của KPDL.

- Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support)

- Điều trị y học (medical treatment): Hiện nay, ứng dụng công nghệ lưu trữ lớn, khai phá dữ liệu trong lĩnh vực y tế để chẩn đoán, phòng ngừa và điều trị bệnh nhằm can thiệp nâng cao sức khỏe con người là hướng nghiên cứu có nhu cầu thực tiễn, được quan tâm tích cực bởi cộng đồng các nhà nghiên cứu. Một số ứng dụng cụ thể của KPDL trong y học:

+ Dự đoán khả năng nhiễm bệnh

+ Dự đoán mức độ nghiêm trọng của virus đối với cơ thể con người

- Text mining & Web mining: KPDL văn bản và KPDL Web là một trong những ứng dụng quan trọng hiện nay. Các bài toán trong KPDL văn bản bao gồm:

+ Tìm kiếm văn bản

+ Phân lớp văn bản

+ Tóm tắt văn bản

+ Phân cụm văn bản

+ Phân cụm các từ mục

+ Đánh chỉ mục các từ tiềm năng

+ Dẫn đường văn bản

Đối với các bài toán trong KPDL Web bao gồm:

+ Thu thập và xử lý dữ liệu Web

+ Phân lớp nhóm các Website có độ uy tín khi truy cập

- Tin sinh học (bio-informatics): KPDL sinh học là một phần rất quan trọng của lĩnh vực Tin-Sinh học (Bioinformatics). Một số ứng dụng của KPDL trong sinh học:

+ Lập chỉ mục, tìm kiếm tương tự, bất thường trong CSDL Gen.

+ Xây dựng mô hình khai phá các mạng di truyền và cấu trúc của Gen, protein

+ Xây dựng các công cụ trực quan trong phân tích dữ liệu di truyền.

- Tài chính và thị trường chứng khoán (finance & stock market): Dữ liệu tài chính trong ngân hàng và trong ngành tài chính nói chung thường đáng tin cậy và có chất lượng cao, tạo điều kiện cho khai phá dữ liệu. Dưới đây là một số ứng dụng điển hình trong khai phá dữ liệu tài chính:

- Dự đoán khả năng vay và thanh toán của khách hàng, phân tích chính sách tín dụng đối với khách hàng.

+ Phân tích hành vi khách hàng (vay, gửi tiền)

+ Phân loại và phân nhóm khách hàng mục tiêu cho tiếp thị tài chính

+ Phát hiện các hoạt động rửa tiền và tội phạm tài chính

- Bảo hiểm (insurance)

- Nhận dạng (pattern recognition)

Trong chương này, luận văn sẽ giới thiệu tổng quan về khai phá dữ liệu bao gồm định nghĩa, một số nghiên cứu, những kỹ thuật khai phá và xử lý dữ liệu hiện nay. Tiếp theo đó là tổng quan về các kỹ thuật khai phá văn bản, ứng dụng trong bài toán phân tích bình luận khách hàng.

1.1.1. Những khái niệm về khai phá dữ liệu

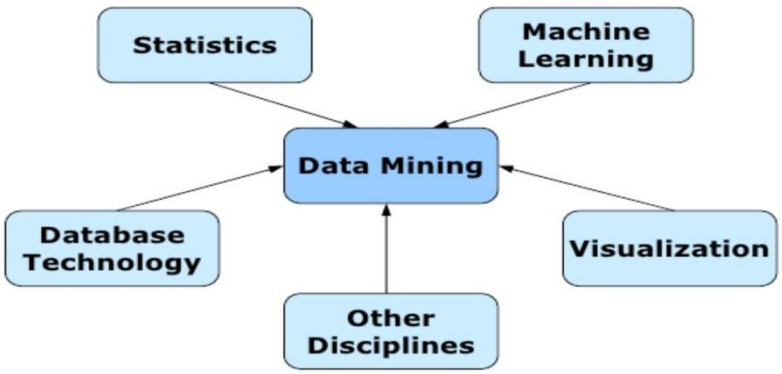
Ngày nay, dữ liệu do con người tạo ra ngày càng nhiều hơn về số lượng, tăng nhanh về khối lượng, phát triển mạnh về quy mô khiến việc phân loại, lựa chọn, khai thác và sử dụng gặp những khó khăn nhất định. Khái niệm khai phá dữ liệu ra đời hỗ trợ những công việc này.

Đến nay, có rất nhiều định nghĩa về khai phá dữ liệu nhưng nhìn chung mỗi định nghĩa đều hướng tới một nhận định. Theo Tom Mitchell [3]: "KPDL là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai." Với một cách tiếp cận ứng dụng hơn, Fayyad [4] đã phát biểu: "KPDL, thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các qui luật, ràng buộc, qui tắc trong cơ sở dữ liệu." Tóm lại, KPDL là một quá trình học tri thức mới từ những dữ liệu đã thu thập được [5,6,7].

Khái niệm về khai phá dữ liệu (Data Mining) hay khám phá tri thức (Knowledge Discovery) có rất nhiều cách diễn đạt khác nhau nhưng về bản chất đó là quá trình tự động trích xuất thông tin có giá trị (Thông tin dự đoán - Predictive Information) ẩn chứa trong khối lượng dữ liệu khổng lồ trong thực tế. Thuật ngữ Data Mining cũng ám chỉ việc tìm kiếm một tập nhỏ có giá trị từ một số lượng lớn các dữ liệu thô.

Hình 1.1 Quá trình trích xuất thông tin có giá trị

Khai phá dữ liệu cũng là một lĩnh vực liên ngành, nơi hội tụ của nhiều học thuyết và công nghệ.



Hình 1.2 Những lĩnh vực liên quan tới khai phá dữ liệu